Hanlong Chen

sl.chen@mail.utoronto.ca | +1 (437) 996-0555 | www.hanlongchen.com

EDUCATION

University of Toronto Sep 2026 - Nov 2030 (expected)
Incoming PhD in Biostatistics

University of Toronto Sep 2024 – Aug 2026
Master of Science in Biostatistics Cumulative GPA: 3.93/4.00
Thesis: Landmark Models for Clustered Data to Account for Within-cluster Correlation of Units
Supervisor: Dr. Aya Mitani

University of Toronto Sep 2019 – Jun 2024
Honours Bachelor of Science in Mathematics and Statistics Graduated with Distinction

RESEARCH EXPERIENCE

University of Toronto, Dalla Lana School of Public Health Toronto, ON
Master Thesis Jan 2025 – Present
Title: Landmark Models for Clustered Data to Account for Within-cluster Correlation of Units
Supervisor: Dr. Aya Mitani

- Established a shared frailty landmark modeling framework incorporating LM2.0 with Gaussian Process-based longitudinal smoothing to address the gap in dynamic prediction for clustered survival data, yielding joint-model-like trajectory reconstruction and improving AUC by 0.15–0.20 over the conventional Cox landmark approach.
- Developed R functions for computationally efficient cluster-adjusted AUC algorithm based on Mann-Whitney formulation, accounting for all four pair-wise censoring configurations (both uncensored, one censored, both censored, tied events), reducing computational complexity from $O(n^2)$ to O(n log n) through stratified counting while avoiding infeasible pairwise comparisons.
- Designed a comprehensive simulation framework generating multilevel longitudinal cluster data (patient, tooth) with time-varying covariates to evaluate landmark model performance under clustered survival settings across cluster sizes, censoring rates, hazard rate and time-varying covariate configurations on the Trillium high-performance computing cluster.

Research Assistant Apr 2025 – Present
Title: Assessing the Reliability of Electronic Dental Records Through Agreement of Probing Pocket Depths Measured Between Two Closely Conducted Periodontal Exams
Supervisor: Dr. Aya Mitani

- Implemented a cumulative link mixed model (CLMM) with multi-level random effects for clustered ordinal clinical data, developed integral-based agreement and Kappa statistics via constrained optimization, and derived a full delta-method variance estimation pipeline for confidence interval construction.

Canadian Primary Care Sentinel Surveillance Network Toronto, ON
Research Assistant Apr 2025 – Present
Title: A Comparison of Sequential and Joint Modeling Approaches for Imputing Longitudinal Laboratory Data to Address Visit-Level Informative Missingness in Electronic Health Records
Supervisor: Dr. Camellia Zakaria & Dr. Aya Mitani

- Analyzed clinically-driven missingness (40%) in longitudinal EMR by stratifying patients by visit density and historical lab-ordering patterns.
- Developed a Sequential FCS-LMM imputation pipeline according to Flux Analysis, using mixed-effects models with random slopes and non-linear time splines, and benchmarked it against MICE-FCS; yields a significant shorter computational time
- Performed multi-method clustering (K-means, Ward's, LCMM) on imputed metabolic profiles, assessing cross-method agreement and imputation-induced stability.
- Characterized distinct metabolic profiles and identified a robust high-risk subtype (elevated HbA1c/TG) consistent across methods, and demonstrated Ward's sensitivity to imputation uncertainty.

Baycrest, Rotman Research Institute Toronto, ON
Graduate Practicum Student Oct 2024 – Present
Title: Cognitive Profiles of People Living with Dementia
Supervisor: Dr. Malcolm Binns

- Implemented PCA-based K-means and Ward's hierarchical clustering on 721 dementia patients, achieving 40% higher R^2 and 75% higher Calinski-Harabasz (CH) index compared to raw score clustering.
- Formulated a standardized evaluation criterion by integrating 9 metrics across stability, interpretability (silhouette score), information retention (cumulative variance), and model fit (CH index) according to the principle of maximum entropy.
- Wrote a machine learning algorithm to optimize PC–k configurations (2–5 PCs × 2–7 clusters) with the developed criterion.
- Identified 4 data-driven distinct cognitive profiles with clinical relevance, revealing age-related patterns ($\rho = 0.36$) differentiating early-onset (EOAD) vs. late-onset dementia (LOAD) characteristics.

University of Toronto, Department of Statistical Sciences Toronto, ON
Undergraduate Research Student Sep 2023 – May 2024
Project: Bayesian Approach for Modelling Inverse Gaussian Observations From a Completely Randomized Design
Supervisor: Dr. Murari Singh

- Derived Jeffreys and GIG-based informative priors for the coefficient of variation under reparameterized Inverse Gaussian and Weibull models, and applied Gibbs sampling in OpenBUGS/R to compare Bayesian 95% HPD intervals with frequentist estimates.

WORK EXPERIENCE

Guangzhou Center for Disease Control and Prevention Guangzhou, China
Research Analyst Intern May – Aug 2021 & May – Aug 2023
Supervisor: Yilie Ma

- Collected, cleaned, and analyzed COVID-19 epidemiological data to identify the impact of vaccination on COVID-19 outbreaks.
- Prepared technical reports supporting epidemic response and public health decision-making.
- Collaborated with cross-functional teams of epidemiologists, clinicians, and data managers to ensure data quality and consistency in a fast-paced public health setting.

Guangzhou Baiyun International Airport Guangzhou, China
Data Analyst Intern Jan 2021 – Apr 2021

- Built a workforce report by analyzing flight volumes, passenger throughput, and multi-scale peak patterns and quantify per-counter passenger processing time and evaluated individual employee on-duty hours and operational efficiency.
- Presented findings to group-level leadership and reasonably justified a need of increasing department's annual personnel budget.

WORKING PAPERS

1. **Hanlong Chen**, Xinyang Feng and Aya Mitani.
   "Landmark Models for Clustered Data to Account for Within-cluster Correlation of Units". *(Manuscript in preparation)*
2. **Hanlong Chen**, Ian Zhang, Bruna Seixas-Lima, Howard Chertkow, ... and Malcolm Binns.
   "Reproducible Dementia Subtyping through Cross-Validated Joint Hyperparameter Optimization of PCA dimensionality and k in Unsupervised Clustering".
3. Aya Mitani, **Hanlong Chen**, Yuning Wang and Christopher McCulloch.
   "Assessing the Reliability of Electronic Dental Records Through Agreement of Probing Pocket Depths Measured Between Two Closely Conducted Periodontal Exams". *(Manuscript in preparation)*
4. **Hanlong Chen**, Bruce Liu, Karim Keshavjee, Aya Mitani, Junzi Chen, Jiakun Lin, and Camellia Zakaria.
   "A Sequential Mixed-Effects Imputation Framework for Visit-Level Informative Missingness and Imputation-Aware Evaluation of Metabolic Subtyping Methods in Longitudinal EHR Data". *(Manuscript in preparation)*

CONFERENCE PRESENTATION

Contributed Talk & Poster Presentation Saskatoon, SK
Statistical Society of Canada (SSC) Annual Meeting & The 13th Annual CSSC May 2025

1. **Hanlong Chen**, Bruna Seixas-Lima, Howard Chertkow and Malcolm Binns.
   "Cognitive Profiles of People Living with Dementia - PCA-Based Clustering Analysis". Modern Approaches for Modelling High-Dimensional and Structured Data Section. *2025 Statistical Society of Canada (SSC) Annual Meeting, p.255*.

AWARD

Graduate Student Research Fellowship Feb 2026
Department of Political Science at the University of Toronto $1,000

Jack and Rita Catherall Fund – Research Grant May 2025
Research Training Centre at Baycrest $500

TEACHING EXPERIENCE

University of Toronto Toronto, ON
Teaching Assistant 2023 – 2026
AI4PH - Handling Missing Data in Health Research; Introduction to Statistical Reasoning and Data Science; Differential Equations I; Introduction to Data Science; Statistics II; Introduction to Applied Statistics; Bayesian Statistical Analysis.

TECHNICAL SKILLS

Programming Languages: R (tidyverse, survival, caret, coxme, glmmTMB, INLA, riskRegression), Python (pandas, numpy, scikit-learn), SAS (Macro, SQL, EG), MATLAB.

Data Tools & Reporting: LaTeX, Git, Linux, SciNet, CDISC (SDTM & ADaM), AI Agents (Claude Code, OpenClaw, Codex).

Statistical Methods: Clinical Trial Design, Survival Analysis, Generalized Linear Mixed Models, Bayesian Methods, Machine Learning (K-means, Hierarchical Clustering, PCA, LCA, LPA, Random Forest, XGBoost, Deep Learning), Time-Series Analysis, Missing Data Imputation, High-dimensional Data Analysis, Simulation Studies.

Languages: English (Advanced), Mandarin (Fluent), Cantonese (Fluent).